Cancer Screening: The Clash of Science and Intuition*

Barnett S. Kramer and Jennifer Miller Croswell

Office of Disease Prevention, Office of the Director, National Institutes of Health, Bethesda, Maryland 20892; email: bk76p@nih.gov, croswellj@od.nih.gov

Annu. Rev. Med. 2009. 60:125–37

The *Annual Review of Medicine* is online at med.annualreviews.org

This article's doi: 10.1146/annurev.med.60.101107.134802

0066-4219/09/0218-0125$20.00

*The views expressed in this article are those of the authors and do not necessarily represent the views of the U.S. federal government or the National Institutes of Health. The U.S. Government has the right to retain a nonexclusive, royalty-free license in and to any copyright covering this paper.

Key Words

five-year survival, selection bias, lead-time bias, length-biased sampling, overdiagnosis

Abstract

The concept of early detection of cancer holds great promise and intuitive appeal. However, powerful biases can mislead clinicians when evaluating the efficacy of screening tests by clinical observation alone. Selection bias, lead-time bias, length-biased sampling, and overdiagnosis are counterintuitive concepts with critical implications for early-detection efforts. This article explains these biases and other common confounders in cancer screening. The most direct and reliable way to avoid being led astray by intuitions is through the use of randomized controlled trials.

INTRODUCTION

The concept of early detection of disease has great intuitive resonance for Americans. It speaks to a sense of individual responsibility and the opportunity to improve one's destiny through action, and sits squarely in line with such aphorisms as "an ounce of prevention is worth a pound of cure" and "a stitch in time saves nine." This idea has informed the practice of medicine for nearly a century. In 1924, the *New York Times* reported a call to public action by Johns Hopkins surgeon Dr. Joseph Colt Bloodgood, who asserted, "Deaths from cancer would be practically eliminated . . . if persons afflicted sought medical aid immediately upon the discovery of a foreign growth in any part of the body" (1). He obviously overestimated the potential impact of available early-detection strategies. Nevertheless, the development and application of new medical technologies have accelerated the actualization of this concept. For example, the American Cancer Society now advertises that with testing, "you have the power to stop colon cancer before it starts" (2). The concept of the worth of screening has taken hold in the American public and in our medical culture. A survey has shown that a substantial proportion feel it is irresponsible for an 80-year-old person to forego cancer screening, and almost three fourths would choose a total-body computerized tomography (CT) scan over a $1000 cash gift (3).

Despite the strength of the messages transmitted to the public about the value of cancer screening, cancer mortality statistics remain sobering. Cancer remains the second most common cause of death in the United States, accounting for 23% of all deaths and dwarfing the third most common cause, stroke, at 6% (see Reference 4, table C). There has been clear progress, but it has been incremental. There has been a decrease of about 10 deaths due to cancer per 100,000 persons per year between 1950 and 2005 (194 versus 184 deaths per 100,000 persons) (5). Population trends reflect a mix of changes in exposures, treatment advances, and screening, so they cannot be used to draw any definitive conclusions about the contribution of early-detection strategies. Nevertheless, the population trends contrast with many public perceptions and clinical intuitions about the magnitude of efficacy of cancer screening.

A core question is, how could Dr. Bloodgood's clinical intuitions and observations have been so misleading? And how can we use scientific methodology to protect us from our strong intuitions? We propose that large-scale randomized trials come to the rescue.

SCREENING VERSUS TREATMENT

Why is it important to look to the highest levels of evidence in making public health policy with respect to cancer screening? After all, most of medical practice is informed by study designs other than randomized trials. Screening for occult disease and treatment of established disease are fundamentally distinct activities, with different risk/benefit ratios to consider. Early-detection activities are aimed at anyone at risk for acquiring the disease of interest; however, the majority of this large pool of individuals will never develop the target cancer and therefore cannot derive the intended benefits of screening (6). However, potential harms associated with the use of screening tests are not contingent on the presence of cancer. It works a bit like a lottery—a few benefit, but downsides are distributed more broadly. As it is difficult to make a well person better off than she already is, the ratio of potential risks to benefits can tip in the wrong direction when screening is applied across a large, generally healthy population (7).

THE ICEBERG MODEL

There is also a tendency to assume that cancers picked up by screening efforts necessarily behave in the same manner as cancers diagnosed by symptomatic presentation. The development of cancer is a lengthy (years or decades) and complex process, the hallmark of which is unrepaired genetic instability leading to distinct, heterogeneous subpopulations of

abnormal cells. As such, cancer can be envisioned as an iceberg of disease, in which the visible tip above the waterline comprises the most aggressive lesions—those that produce symptoms and clinical disease. The majority of our body of knowledge concerning the natural history of malignancies comes from observations of these symptomatic lesions. Underneath the water's surface, however, there are multiple subpopulations of cells, ranging from those with genetic and epigenetic changes to those with phenotypic abnormalities. Some of these subpopulations will look like typical cancers to a trained pathologist. However, a static snapshot may not reflect dynamic cellular behavior. Early-detection methods, by definition, attempt to dip below the waterline and pick up silent lesions; but the natural history of these asymptomatic lesions has not been observed and is essentially unknown (8). In fact, because of the geometry of the iceberg, even a modestly sensitive screening test may detect more cancers whose natural history is not known than whose natural history is known.

INCIDENCE, MORTALITY, AND FIVE-YEAR SURVIVAL RATES

What has been observed is that the more we search for cancer, the more cancer we find. To illustrate this point, **Figure 1*a*** shows the age-adjusted incidence rates between 1973 and 2003 for four cancers—prostate, breast, melanoma, and thyroid—that were targeted for early-detection efforts during the time interval. These rates apply to U.S. citizens ages 40 and older and were obtained from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) database (9).

Evidence suggests that a large proportion of the rise in incidence in each cancer can be accounted for by new cases of localized disease. **Figure 1*a*** also has the rates of localized and distal lesions (as seen at time of diagnosis) superimposed on the overall incidence rates (9). In each case, trends for localized disease roughly track overall incidence trends, whereas advanced disease trends are more stable. It is the latter that have the greatest lethal potential.

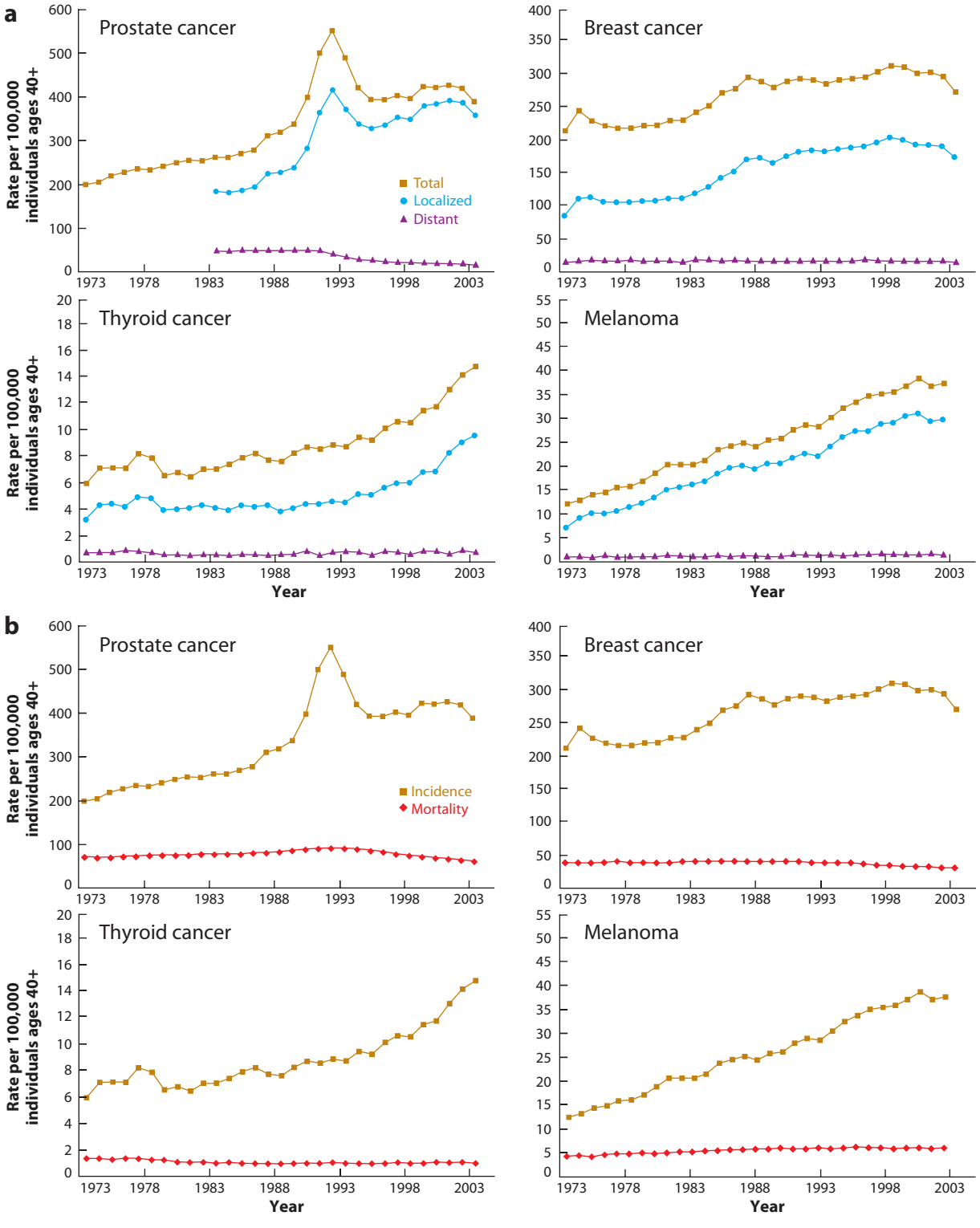
If screening and early detection were simply pulling late-stage cancers out of the future, then, over time, one would expect a resulting decrease in advanced cancers and hope for their eventual disappearance. This has not been clearly observed. The implication is that early screening can detect large numbers of cancers that were not destined to metastasize. It is difficult to attribute these trends to widespread exposures to new, unknown carcinogens because most carcinogens should cause cancers of all stages—not just early-stage ones. A practical extension of these concepts is that the goal of cancer screening is not to detect as many cancers (or even early-stage cancers) as possible but to decrease mortality.

Figure 1*b* superimposes the incidence and mortality rates for prostate, breast, melanoma, and thyroid cancers (9). In every case, there is a disparity between the number of new cases diagnosed and the number of cancer deaths averted. For prostate cancer, when compared to the 1973 (pre–screening era) baseline, at the 1992 peak incidence, there was an excess of 352 new cases per 100,000 persons diagnosed. However, the lowest yearly mortality rate (in 2003) was 13 fewer deaths per 100,000 compared with the baseline. For breast cancer, the peak difference in yearly incidence to baseline was 52 cases per 100,000 women; the peak difference in yearly mortality to baseline was 8 deaths per 100,000 women. Thyroid and melanoma have similar findings.

One cannot utilize ecological data to draw definitive conclusions about the effectiveness of a screening modality, but these are counterintuitive findings that can generate cognitive dissonance. After all, the media frequently report on increasing five-year survival rates for a wide range of cancer types as a measure of progress. Indeed, using our four examples, there have been notable increases in survival trends over the past 30 years. Comparing the time intervals of 1975–1979 and 1996–2003, as reported by the SEER database, overall five-year survival rates increased substantially for all four cancer

Mortality: the total number of deaths from a given cause over a defined time interval (e.g., 1 year) divided by the number of persons at risk for the disease in the population during the same interval

Five-year survival: the number of persons diagnosed with the condition of interest who are alive five years after diagnosis, divided by the total number of persons with the disease during those five years



types. How can mortality and survival provide such disparate pictures of progress?

Although it is easy to conflate the two concepts mentally, survival and mortality rates are distinct measures and are calculated in different ways. Mortality is defined as the total number of deaths from a given cause over a defined time interval (e.g., one year) divided by the number of persons at risk for the disease in the population during the same interval. Survival, on the other hand, is the number of persons diagnosed with the condition of interest who are alive a defined number of years after diagnosis (e.g., five years), divided by the total number of persons with the disease during that time interval. The distinction between the denominators points to one of the critical issues that can lead a clinician astray in judging the worth of a screening test: The clinician is only observing the restricted denominator of diagnosed patients.

BIASES IN SCREENING

Important biases that are frequently associated with observations (and observational studies) of screening tests can confound assessment of screening-test efficacy. This section explains four of these biases.

Lead-Time Bias

Early detection advances what would have been the original date of diagnosis to an earlier point in time, but it does not necessarily follow that the patient's time of death will be delayed. For example, if a particular disease has no known treatment, earlier detection can have no impact on the lifespan of an affected person. In **Figure 2*a***, the two lines represent two lifespans. In the one case, cancer is detected through symptoms, and the person dies at a set point. In the second case, the cancer is detected while asymptomatic through screening. The proportion of the lifespan affected by disease has been extended—that is, the person is a patient longer—but total years of life remain exactly the same.

Lead-time bias: bias in survival rates caused by the time interval between diagnosis at the asymptomatic stage (by screening) and the onset of clinical symptoms. Early detection, by advancing the date of diagnosis, necessarily adds apparent survival time compared with symptomatic detection; however, this may not translate into a longer overall lifespan for a given individual

Lead-time bias is a mathematical certainty with all early detection and has important implications for the use of five-year survival as an indicator of screening effectiveness. To see why this is true, consider the following hypothetical example: Suppose there is a cancer that kills 100% of people within four years after the onset of symptoms (i.e., four years after clinical diagnosis). The five-year survival rate is 0%. Now suppose we introduce a new screening test, which detects the cancer five years before symptoms begin. The five-year survival rate from date of diagnosis is now 100% for this screened population, even though we have done nothing to alter the course of disease! In fact, if a given therapy for a screen-detected cancer were toxic enough to shorten life expectancy on average, the screening test will still appear to improve survival if the decrease is smaller than the lead-time.

Conflicting results between five-year survival and cause-specific mortality rates have been documented in clinical trials. One example is the Mayo Lung Project, a randomized controlled trial of lung cancer screening that compared chest X-ray plus sputum cytology versus usual care in male smokers (14, 15).

Figure 1

Cancer rates from the database of the Surveillance, Epidemiology, and End Results (SEER) Program (**http://www.seer.cancer.gov**) for cancers of the prostate, breast, thyroid, and skin (melanoma) per 100,000 persons ages 40 and older in the United States from 1973 through 2003. Rates were obtained by running the following commands: SEER*Stat Database: Incidence—SEER 9 Regs Limited-Use, Nov 2006Sub (1973–2004)—Linked To County Attributes—U.S., 1969–2004 Counties and SEER*Stat Database: Mortality—All COD, Aggregated With State, U.S. (1969–2004). (*a*) Total, localized, and distant cancer incidence rates. As defined by SEER historic stage A, "localized" refers to an invasive cancer confined to the organ of origin; "distant" describes metastasis to organs not adjacent to the organ of origin or to distal lymph nodes. (b) Total incidence and mortality rates. From Reference 9. Underlying mortality data provided by NCHS (**http://www.cdc.gov/nchs**). These concepts are discussed elsewhere in greater detail (10–13).

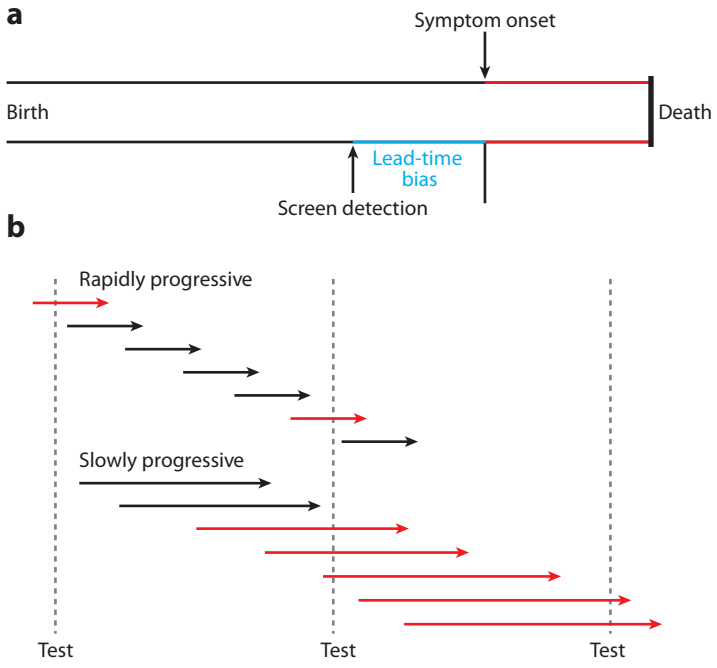


Figure 2

(*a*) Lead-time bias. Early detection necessarily advances the date of diagnosis of a cancer compared with clinical detection; however, in this case, although the individual lives longer with a diagnosis of cancer, there is no change in the date of death. Adapted from Reference 38 with permission. (*b*) Length-biased sampling. The arrows represent tumors. The body of the arrow represents the preclinical growth period; the arrowhead represents the onset of symptomatic disease. The vertical dotted lines represent application of a screening test. Screening is more effective at detecting slow-growing, less aggressive tumors, because these lesions have a longer preclinical period during which they can be detected by tests. Adapted from Reference 38a with permission.

Five years after diagnosis, lung cancer survival was 35.6% (95% CI, 28.6%–42.6%) among participants in the chest X-ray arm versus 18.5% (95% CI, 11.5%–25.5%) in the control arm. Mortality rates, however, did not differ statistically significantly between the two groups and even trended towards an *increase* in deaths among screened individuals: 4.4 deaths per 1000 person-years for screened individuals (95% CI, 3.9–4.9) and 3.9 deaths per 1000 person-years for the control group (95% CI, 3.5–4.4) (14). An analysis of SEER data from 1950 to 1995 demonstrated no correlation between five-year survival and mortality rates for 20 types of cancer during this time interval (Pearson r = 0.00) (16). Comparisons of survival when evaluating screening tests can be highly misleading.

Healthy-volunteer bias: trial selection bias due to fundamental differences between individuals willing to participate in screening efforts and those who are not

Healthy-Volunteer Bias

A second common phenomenon is healthy-volunteer bias. There are usually fundamental differences between those individuals willing to participate in early-detection efforts and those who are not. Volunteers may be more attuned to health messages and more apt to adhere to health providers' recommendations. They may also be from a higher socioeconomic group and have better access to quality healthcare. As an example, a group of Japanese researchers surveyed individuals who chose to participate in a nationally offered mass screening program versus those who chose not to. They found that screening participants consumed more vegetables, seaweed, milk, and dietary fiber, and were less likely to be smokers, than those individuals who chose not to be screened (17).

Follow-up data from a British cohort study of asymptomatic women who volunteered to receive screening for ovarian cancer revealed that these women had lower than expected mortality rates from colorectal, stomach, lung, and cervical cancer, even though transvaginal ultrasound would have no impact on these tumors (18). Investigators with the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized cancer screening trial documented this phenomenon on a broader scale. Participants in both screening and usual-care arms had lower than expected mortality rates for cardiovascular, respiratory, and digestive diseases; diabetes; all nonprostate, lung, colorectal, or ovarian cancers; and even injuries and poisonings (19). There are clearly important confounders that track with populations interested in early-detection methods. These qualities can amplify or create an apparent benefit even if there is no mortality reduction.

Length-Biased Sampling

Early diagnostic tools are more likely to pick up less aggressive lesions than rapidly lethal

lesions. Use of such tools automatically appears to be associated with better outcomes through this unequal sampling of tumors. **Figure 2*b*** illustrates why length-biased sampling occurs: The slow-growing, less aggressive cancers have a longer preclinical period than more rapidly fatal cancers. As such, the screening tool has a greater likelihood of detecting these lesions than tumors that grow and spread quickly. Attempts to directly compare outcomes between screen-detected and symptom-detected cancers are thus inherently biased. An investigation of a mass CT screening program for lung cancer in Japan generated evidence to support this concept. The majority of screen-detected lung cancers were well-differentiated adenocarcinomas that had substantially extended doubling times when compared to the behavior of symptomatic tumors monitored by chest X-ray (mean of 452 versus 164 days) (20, 21).

Overdiagnosis

An extreme form of length-biased sampling is overdiagnosis. It occurs when a pathologist makes a histological diagnosis of cancer, but the lesion lacks true malignant potential despite its appearance (sometimes called pseudodisease) or is so slow-growing that the individual will die from a different, competing cause before her health is threatened by the lesion. Cancer is primarily a disease of aging, so competing causes of death can account for a large proportion of deaths even in people who coincidentally have indolent cancers. For those who have witnessed the devastating effects of cancer on an individual, overdiagnosis is highly counterintuitive. The concept implies that there are lesions that can be detected by the same screening modalities that identify "true" cancers, and are indistinguishable under the microscope from "true" cancer, but do not behave like "true" cancer.

A growing body of evidence suggests that the appearance of overdiagnosis is a common price of early-detection efforts, even for effective screening tests. Autopsy studies provided the first clues of pseudodisease. For example, combined data from seven autopsy series of women with no prior history of breast cancer demonstrated a median prevalence of undetected ductal carcinoma in situ of 9% (ranging to 15%) (22). A 1997 review of eight autopsy studies found that nearly one third of asymptomatic men 50 years and older who had died from other causes had histologically confirmed prostate cancer; 43% of asymptomatic men 80 years and older harbored occult lesions (23). Similarly, investigators in Finland found a 36% prevalence of undetected papillary carcinoma of the thyroid (24). These reports lend weight to the previously introduced analogy of cancer as an iceberg of disease. There appears to be a large reservoir of "cancers" in the population that could be detected if we looked for them with screening tests.

Evidence exists that some asymptomatic cancers have very different natural histories from what we would expect on the basis of observations of symptomatic tumors. Several studies have investigated the impact of a Japanese national infant screening program for neuroblastoma. The incidence rates rose sharply with intensive screening efforts. One cohort study in the Saitama Prefecture found that between 1981 and 1992, incidence rose from a baseline of 33–69 cases per million infants under one year of age to a high of 260 cases per million (25). The majority of the increase was attributable to early-stage (I/II) diagnoses. However, across the studies, the rates of advanced-stage (III/IV) disease and deaths due to neuroblastoma remained the same (25–27). Mass neuroblastoma screening was abandoned, and in interpreting these unexpected findings, the investigators hypothesized that the tumors identified by screening must be spontaneously regressing abnormalities—a fundamentally different form of disease from the symptomatic variety (25, 26).

A three-year study of annual spiral CT scans for the early detection of lung cancer generated an equally intriguing finding: Rates of lung cancer detection among smokers and never-smokers were nearly identical (0.46% versus 0.41%) (28). In fact, the first scan picked up 13 lesions in the nonsmoking group and only

Length-biased sampling: early diagnostic tools disproportionately detect slower-growing cancers compared with cancers discovered as the result of symptoms

Overdiagnosis: an extreme form of length-biased sampling, in which, despite its pathological appearance, the cancer either has no malignant potential ("pseudodisease") or is so indolent that it cannot alter remaining lifespan because the person will die of another cause first

10 in the smokers! Considering that 85% of the attributable risk of developing lung cancer has been ascribed to smoking (29), these results raise serious questions about the potential for overdiagnosis of "cancers" that would never have come to attention had it not been for a screening CT.

If a person has a favorable outcome, there is no way of knowing whether he or she lived because of successful therapy or because the lesion was actually of no lethal potential (30). However, one indicator that raises the suspicion of overdiagnosis is when there is a persistent imbalance in the overall cancer incidence rates in the screened and unscreened arms, along with a persistent imbalance in the proportion of early- versus late-stage disease diagnosed, but relative equivalence between the groups' mortality rates (13).

Extended follow-up of the Mayo Lung Project revealed this exact situation. As discussed above, there were no observed differences in lung cancer mortality between arms. At the end of this trial, investigators noted an excess of 46 cases of lung cancer in the screened group versus controls; 16 years after the conclusion of the trial, the imbalance persisted. Furthermore, there were notable differences in the number of early- versus late-stage tumors detected in the two groups. As might be anticipated, the screened arm had a greater number of early-stage lung cancer cases (99 versus 51 cases). The number of advanced tumors diagnosed, however, remained essentially identical (107 versus 109 cases) (31). Because more tumors were found in the screened group, the stage-specific treatments provided for identified lesions were identical in both groups, and the same number of persons died from lung cancer in each arm, the researchers concluded that the excess lesions could have posed little true threat to health (31, 32). A 2008 review of lung cancer screening trials estimated that at least 25% of tumors identified by chest X-ray appear to be instances of overdiagnosis (33).

Recent estimations of overdiagnosis in breast cancer screening have also been published. Zackrisson et al. (34) performed a 15-year follow-up study of the Malmö mammographic screening trial. These investigators calculated a 10% rate of overdiagnosis among the screened women; this calculation was deemed an underestimate in an accompanying letter to the editor, which proposed a revised rate of 25% (35). A Cochrane review of mammography trials calculated the overdiagnosis rate at 32% (36).

Overdiagnosis is likely to accompany a broad range of, if not all, cancer screening tests (**Figure 1**). The reason overdiagnosis is of such concern is that an individual with pseudodisease cannot, by definition, benefit from treatment, but he or she can experience morbidity or even death from therapy.

A MISLEADING POSITIVE FEEDBACK LOOP

The potential benefits of cancer screening are intuitively powerful and have been broadly advertised by public health and advocacy organizations. The concept that screening might also have the potential to negatively affect health, however, is infrequently part of health-promotion messages, many of which have soundbite quality. One reason for this disconnect may be that at the level of individual experience and observation, getting screened is a behavior that almost invariably receives positive reinforcement, irrespective of the actual outcome. If the test is positive, both patient and clinician feel gratified that the disease was "caught early"; if negative, the exam provides reassurance and a sense one is being responsible for one's health. If the diagnostic workup or resulting treatment cause complications (for example, a pneumothorax caused by lung biopsy), the clinician and patient will likely consider such effects a small price to pay for survival (37). This positive feedback loop is further reinforced by the apparent spike in incidence that accompanies the use of a new screening modality, coupled with a greater proportion of diagnoses of milder forms of disease. As a result, there can be a false inflation of the relative

impact of the disease on the population compared with other illnesses (38). Superimposed on this is the real or perceived legal jeopardy of physicians if an advanced cancer is diagnosed in a patient who was not encouraged to have a screening test (39, 40).

ASSESSING THE BENEFITS AND HARMS OF EARLY DETECTION

Diagnostic Follow-Up

Screening tests are, in general, quick, relatively noninvasive procedures that cause minor discomfort. One of the World Health Organization's ten fundamental screening principles is that "the test should be acceptable to the population"—otherwise, the population will simply refuse to undergo it (41). However, follow-up testing to confirm a positive screen is not always innocuous, and any adverse events resulting from these more invasive tests must be considered part of the clinical cascade associated with early-detection efforts.

For example, screening colonoscopy (a simultaneous screening and diagnostic modality) is associated with known complications, the most important of which are colonic bleeding and perforation. A recent cohort study reported an overall rate of 7 serious complications per 1000 diagnostic colonoscopies (that is, colonoscopy coupled with accompanying biopsy or polypectomy). In this population, there were 1.1 perforations, 4.8 serious bleeding episodes, and 0.6 deaths per 1000 diagnostic colonoscopies (42). Baseline results for ovarian cancer screening using transvaginal ultrasound and CA-125 in the PLCO cancer screening trial revealed that 570 women screened underwent laparotomy or laparoscopy to follow-up a positive screening exam; 29 tumors were diagnosed. This means 32% of women with a positive screening test underwent a major surgical procedure but did not have cancer, and 95% of all abdominal/pelvic surgeries were for false-positive tests (43).

False-Positive Results

False-positive screening exams cause anxiety that may persist beyond resolution of the test result. A 2007 systematic review of the long-term effects of false-positive mammograms found that, compared with women who had received normal results, women with false-positive test results utilized mental healthcare professionals more frequently and had higher levels of anxiety, apprehension, and intrusive thoughts specific to fears about breast cancer (44).

It is a mathematical certainty that any screening test with a specificity of less than 100% will generate an increasing burden of false-positive results with repeated applications of the test. For a woman without breast cancer between the ages of 40 and 69, who receives annual screenings, the estimated 10-year risk of a false-positive mammogram is ~49%; the estimated 10-year risk of a false-positive breast biopsy is 19% (45). The cumulative risk for an individual to receive at least one positive test for a multiple-modality screening program over the course of 14 tests (three years) has been projected to be about 60% for men and 50% for women (46).

Treatment Regimens

The benefits and harms of attendant treatment regimens also influence the ultimate utility of a given screening modality. The success of a screening test necessarily depends on the availability of effective therapies for the target lesion. If there is no known treatment for a disease, there is nothing to be gained by detecting it sooner. The overall risk/benefit ratio for early detection of a given cancer across a population also depends on the potential harms of therapy, particularly in relation to the degree of overdiagnosis associated with a specific screening modality.

Cancer therapies often have short- and long-term morbidities; in some cases, the treatments also carry a mortality risk. For example, a 2008 study of post-treatment quality of life

among survivors of localized prostate cancer who chose radical prostatectomy, brachytherapy, or external-beam radiotherapy found that at one year after therapy, 75%, 64%, and 54% of men, respectively, no longer experienced erections firm enough for intercourse; 16%, 8%, and 6% leaked urine more than once a day; and 3%, 14%, and 10% experienced bowel urgency described as "a moderate or big problem" (47). Similar findings were described in the Scandinavian Prostatic Cancer Group Study, and symptom prevalence persisted five years post-therapy in the Prostate Cancer Outcomes Study (48, 49). In-hospital or 30-day mortality rates for Medicare patients treated with surgery for lung cancer have been documented to range from 4%–6% for a lobectomy to 11%–16% for a pneumonectomy (50). For cases of overdiagnosis, in which no benefits can be reaped by the individual, these morbidity and mortality risks are of considerable concern.

SOLUTIONS: PROTECTING OURSELVES FROM OUR INTUITIONS

The concept of early detection has great promise embedded within it. The opportunity to act early and derail the progression of a lesion, and thereby preempt death, is an attractive one. However, as we have seen, screening also has embedded within it a host of confounding and complicating factors that can be difficult, if not outright impossible, to detect through personal experience and clinical observation alone. How then do we disentangle a cancer screening test's true efficacy, and how do we appropriately weigh the benefits and harms of that early-detection effort? As Sir Richard Peto has succinctly concluded, "There is simply no serious scientific alternative to the generation of large-scale randomized evidence" (51).

Randomized clinical trials are the most direct and reliable way to ensure that all of the potential biases discussed above are controlled for. Personal experiences with a screening tool are insufficient; high-quality clinical trial evidence must be used to make decisions that can potentially affect large proportions of an entire population. Constructing an analytic framework, such as those employed regularly by the U.S. Preventive Services Task Force, can also help to avoid the mental shortcuts and assumptions that can easily lead one astray in cancer screening (52).

Figure 3 provides a generic example of a screening analytic framework. It explicitly lays out each step used to assess the sum total of risks and benefits associated with a given screening test. It reminds the user that there are benefits and harms to consider from the screening modality itself, as well as from the resulting treatments. It draws a visible distinction between intermediate or surrogate outcomes (such as a change in a lab value) and the final outcome of interest—a reduction in morbidity

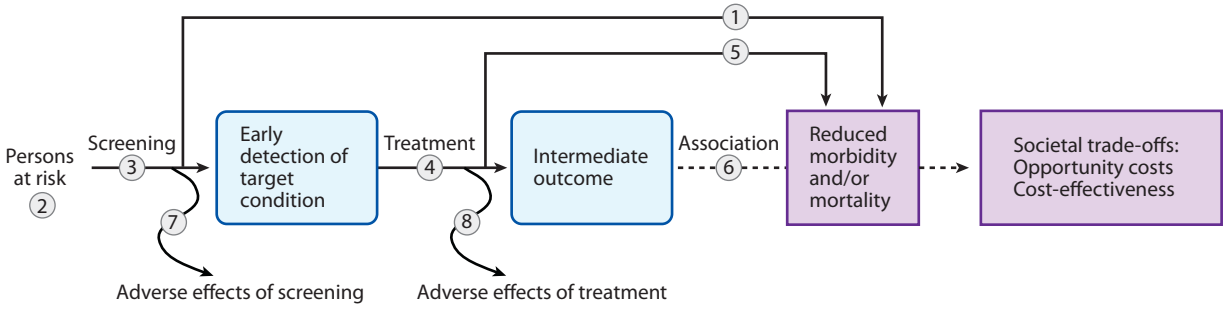


Figure 3

Sample analytic framework for evaluating a screening test, adapted from the U.S. Preventive Services Task Force. Adapted from Reference 52 with permission.

and/or mortality. Finally, the framework highlights the distinction between direct (Path 1 in **Figure 3**) and indirect evidence of efficacy. When evaluating studies of screening, it is helpful to consider where the data fit within this causal pathway: What part of the overarching question does it actually address, and what parts are ignored? This exercise can remind us of any remaining gaps in the evidence chain that need to be filled in before we can draw a reliable conclusion about the screening modality.

Although clinical intuition is a fundamental "art" of medicine, in the field of cancer screening it is easy to be misled. Powerful, pervasive biases make reliance on experience alone a dangerous strategy. Successful evaluation of early-detection efforts requires strict adherence to the scientific method to protect us from simply ratifying our desires. As Roman playwright Terence noted, "One easily believes what one earnestly hopes for." We should harness this passion to generate evidence as strong as our messages. At the very least, we should be aware that soundbites can do injustice to complex trade-offs when proposing cancer screening tests to a healthy population.

SUMMARY POINTS

1. Powerful biases operate in screening tests: selection bias, lead-time bias, length-biased sampling, and overdiagnosis.
2. Even a harmful screening test can appear to be beneficial when evaluated by observation alone.
3. Randomized controlled trials are the most reliable and direct way to ensure the efficacy of cancer screening tests.

DISCLOSURE STATEMENT

Barnett Kramer, as part of his official duties for the U.S. government, serves in leadership positions for two large randomized controlled cancer screening trials funded by the U.S. National Cancer Institute: (*a*) the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial and (*b*) the National Lung Screening Trial (NLST). Jennifer Miller Croswell is not aware of any factors that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. 1924. Cure for cancer in prompt action. *NY Times*, Jun. 7, p. 25
2. American Cancer Society. 2008. *They know how to prevent colon cancer—and you can, too*. **http://americancancersociety.org/docroot/PED/content/PED_5_1x_They_Know_How_to_Prevent_Colon_Cancer.asp?from=colontesting**
3. Schwartz LM, Woloshin S, Fowler FJ, et al. 2004. Enthusiasm for cancer screening in the United States. *JAMA* 291(1):71–78
4. Kung HC, Hoyert DL, Xu JQ, et al. 2008. *National vital statistics reports. Deaths: final data for 2005*. Vol. 56, No. 10. **http://www.cdc.gov/nchs/data/nvsr/nvsr56/nvsr56_10.pdf**
5. American Cancer Society. 2008. *Cancer statistics: a presentation from the American Cancer Society*. **http://www.cancer.org/downloads/STT/Cancer_Statistics_2008.ppt#399,4,Change in the US Death Rates* by Cause, 1950 & 2005**
6. Kramer BS. 2004. The science of early detection. *Urol. Oncol.* 22:344–47
7. Miller AB. 2006. Design of cancer screening trials/randomized trials for evaluation of cancer screening. *World J. Surg.* 30:1152–62
8. Kramer BS, Brawley OW. 2000. Cancer screening. *Hematol. Oncol. Clin. North Am.* 14(4):831–48

9. National Cancer Institute DCCPS, Surveillance Research Program, Cancer Statistics Branch. 2007. Surveillance, Epidemiology, and End Results (SEER) Program. **http://www.seer.cancer.gov**
10. Welch HG. 2004. *Should I Be Tested for Cancer? Maybe Not and Here's Why*, pp. 61–62. Berkeley: Univ. Calif. Press
11. Anderson WF, Jatoi I, Devesa SS. 2006. Assessing the impact of screening mammography: breast cancer incidence and mortality rates in Connecticut (1943–2002). *Breast Cancer Res. Treat.* 99(3):333–40
12. Davies L, Welch HG. 2006. Increasing incidence of thyroid cancer in the United States, 1973–2002. *JAMA* 295(18):2164–67
13. Welch HG, Woloshin S, Schwartz LM. 2005. Skin biopsy rates and incidence of melanoma: population based ecological study. *BMJ* 331:481
14. Marcus PM, Bergstralh EJ, Fagerstrom RM, et al. 2000. Lung cancer mortality in the Mayo Lung Project: impact of extended follow-up. *J. Natl. Cancer Inst.* 92(16):1308–16
15. Fontana RS, Sanderson DR, Woolner LB, et al. 1991. Screening for lung cancer: a critique of the Mayo Lung Project. *Cancer* 67(Suppl.):1155–64
16. Welch HG, Schwartz LM, Woloshin S. 2000. Are increasing 5-year survival rates evidence of success against cancer? *JAMA* 283(22):2975–78
17. Suzuki K, Nakaji S, Tokunaga S, et al. 2005. Confounding by dietary factors in case-control studies on the efficacy of cancer screening in Japan. *Eur. J. Epidemiol.* 20:73–78
18. Crayford TJ, Campbell S, Bourne TH, et al. 2000. Benign ovarian cysts and ovarian cancer: a cohort study with implications for screening. *Lancet* 355:1060–63
19. Pinsky PF, Miller A, Kramer BS, et al. 2007. Evidence of a healthy volunteer effect in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. *Am. J. Epidemiol.* 165(8):874–81
20. Hasegawa M, Sone S, Takashima S, et al. 2000. Growth rate of small lung cancers detected on mass CT screening. *Br. J. Radiol.* 73:1252–59
21. Usuda K, Saito Y, Sagawa M, et al. 1994. Tumor doubling time and prognostic assessment of patients with primary lung cancer. *Cancer* 74:2239–44
22. Welch HG, Black WC. 1997. Using autopsy series to estimate the disease "reservoir" for ductal carcinoma in situ of the breast: How much more breast cancer can we find? *Ann. Intern. Med.* 127(11):1023–28
23. Coley CM, Barry MJ, Fleming C, et al. 1997. Early detection of prostate cancer. Part I: prior probability and effectiveness of tests. The American College of Physicians. *Ann. Intern. Med.* 126(5):394–406
24. Harach HR, Franssila KO, Wasenius VM. 1985. Occult papillary thyroid carcinoma of the thyroid: a "normal" finding in Finland. *Cancer* 56:531–38
25. Yamamoto K, Hayashi Y, Hanada R, et al. 1995. Mass screening and age-specific incidence of neuroblastoma in Saitama Prefecture, Japan. *J. Clin. Oncol.* 13(8):2033–38
26. Honjo S, Doran HE, Stiller CA, et al. 2003. Neuroblastoma trends in Osaka, Japan, and Great Britain 1970–1994, in relation to screening. *Int. J. Cancer* 103(4):538–43
27. Ajiki W, Tsukuma H, Oshima A, et al. 1998. Effects of mass screening for neuroblastoma on incidence, mortality, and survival rates in Osaka, Japan. *Cancer Causes Control* 9(6):631–36
28. Sone S, Li F, Yang Z-G, et al. 2001. Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner. *Br. J. Cancer* 84(1):25–32
29. Chyou PH, Nomura AM, Stemmermann GN. 1992. A prospective study of the attributable risk of cancer due to cigarette smoking. *Am. J. Public Health* 82(1):37–40
30. Black WC. 2000. Overdiagnosis: an underrecognized cause of confusion and harm in cancer screening. *J. Natl. Cancer Inst.* 92(16):1280–82
31. Marcus PM, Bergstralh EJ, Zweig MH, et al. 2006. Extended lung cancer incidence follow-up in the Mayo Lung Project and overdiagnosis. *J. Natl. Cancer Inst.* 98(11):748–56
32. Wolpaw DR. 1996. Early detection in lung cancer: case finding and screening. *Med. Clin. North Am.* 80(1):63–82
33. Reich JM. 2008. A critical appraisal of overdiagnosis: estimates of its magnitude and implications for lung cancer screening. *Thorax* 63:377–83
34. Zackrisson S, Andersson I, Janzon L, et al. 2006. Rate of overdiagnosis of breast cancer 15 years after end of Malmö mammographic screening trial: follow-up study. *BMJ* 332:689–92

35. Welch HG, Schwartz LM, Woloshin S. 2006. Ramifications of screening for breast cancer: 1 in 4 cancers detected by mammography are pseudocancers. *BMJ* 332:727
36. Gøtzsche PC, Nielsen M. 2006. Screening for breast cancer with mammography. *Cochrane Database Syst. Rev.* 4:CD001877. DOI: 10.1002/14651858.CD001877.pub2
37. Ransohoff DF, Collins MM, Fowler FJ. 2002. Why is prostate cancer screening so common when the evidence is so uncertain? A system without negative feedback. *Am. J. Med.* 113:663–67
38. Black WC, Welch HG. 1997. Screening for disease. *Am. J. Roentgenol.* 168:3–11
38a. Black WC, Welch HG. 1998. Advances in diagnostic imaging and overestimations of disease prevalence and the benefits of therapy. *N. Engl. J. Med.* 238:1237–43
39. Merenstein D. 2004. Winners and losers. *JAMA* 291(1):15–16
40. Krist AH, Woolf SH, Johnson RE. 2007. How physicians approach prostate cancer screening before and after losing a lawsuit. *Ann. Fam. Med.* 5(2):120–25
41. Wilson JMG, Jungner G. 1968. *Principles and practice of screening for disease*. Geneva: WHO. **http://whqlibdoc.who.int/php/WHO_PHP_34.pdf**
42. Levin TR, Zhao W, Conell C, et al. 2006. Complications of colonoscopy in an integrated health care delivery system. *Ann. Intern. Med.* 145:880–86
43. Buys SS, Partridge E, Greene MH, et al. 2005. Ovarian cancer screening in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial: findings from the initial screen of a randomized trial. *Am. J. Obstet. Gynecol.* 193:1630–39
44. Brewer NT, Salz T, Lillie SE. 2007. Systematic review: the long-term effects of false-positive mammograms. *Ann. Intern. Med.* 146:502–10
45. Elmore JG, Barton MB, Moceri VM, et al. 1998. Ten-year risk of false positive screening mammograms and clinical breast examinations. *N. Engl. J. Med.* 338(16):1089–96
46. Miller JH, Kramer BS, Kreimer A, et al. 2007. Cumulative false positives (FP) in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. *J. Clin. Oncol.* 25(18):61s (Abstr.)
47. Sanda MG, Dunn RL, Michalski J, et al. 2008. Quality of life and satisfaction with outcome among prostate-cancer survivors. *N. Engl. J. Med.* 358(12):1250–61
48. Steineck G, Helgesen F, Adolfsson J, et al. 2002. Quality of life after radical prostatectomy or watchful waiting. *N. Engl. J. Med.* 347(11):790–96
49. Potosky AL, David WW, Hoffman RM, et al. 2004. Five-year outcomes after prostatectomy or radiotherapy for prostate cancer: the prostate cancer outcomes study. *J. Natl. Cancer Inst.* 96(18):1358–67
50. Birkmeyer JD, Siewers AE, Finlayson EVA, et al. 2002. Hospital volume and surgical mortality in the United States. *N. Engl. J. Med.* 346(15):1128–37
51. Peto R, Collins R, Gray R. 1995. Large-scale randomized evidence: large, simple trials and overviews of trials. *J. Clin. Epidemiol.* 48(1):23–40
52. Harris RP, Helfand M, Woolf SH, et al. 2001. Current methods of the U.S. Preventive Services Task Force: a review of the process. *Am. J. Prev. Med.* 20(3S):21–35

Annual Review of Medicine

Volume 60, 2009

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Medicine* articles may be found at http://med.annualreviews.org/errata.shtml